Gradient Harmonicity in Compounds

Péter Rebrus and Miklós Törkenczy
Research Institute for Linguistics, Hungarian Academy of Sciences (MTA)

1 Introduction: morphological and phonological domains

Some phonological phenomena only obtain within a morphological circumscribed domain. In a strict interpretation this means that morphological parsing *unambiguously* identifies the phonological domain within which the phonological phenomenon applies. There are two problems with this interpretation. First, there are morphology-phonology mismatches in many languages. In these cases, the phonological and morphological domains are misaligned with respect to a domain-sensitive phonological phenomenon so that the boundaries of these domains do not coincide: a morphologically complex form may behave as a single domain phonologically ("underparsing") or a single morphological domain may be more than one domain phonologically ("overparsing"). Traditionally, mismatches are considered to be exceptional and are handled by *ad hoc* devices, i.e. these models establish no connection between the behaviour of these forms and any other property they may have in a systematic and explicit way.

Furthermore, the strict interpretation assumes that morphological parsing is *categorical* in the sense that the morphologically determined boundaries that potentially partition a form are either present in a representation which is to be interpreted phonologically or not, i.e. a given form is either complex or simplex. This holds true independently of (i) how domains are encoded into representations, by segment-like boundary-markers (as in the SPE) or structurally (as in autosegmental and/or lexical models of phonology, e.g. McCarthy 1981, Kiparsky 1982), and (ii) whether we use one or more than one type of boundary marker since in either case a given boundary marker is either present to partition a form or not. Using more than one type of boundary makes it possible to express some degree of graduality but even in this case boundary marking remains categorical (since the different types of boundaries are either present or absent and, accordingly, the form is either complex or simplex) and arbitrary since there is no systematic connection between the presence/absence and the types of these boundaries and any property of the forms in question other than the domain-sensitive phonological phenomenon whose application requires the boundaries. Phonological variation related to morphological structure is especially problematic for the traditional view. A representational approach can analyse phonological vacillation resulting from mismatch in two ways: (a) we can assume multiple underlying morphological and phonological representations for the relevant form, one which is internally divided into domains and another that is undivided, and derive the phonological vacillation the form displays from this parallelism; or (b) we can assume that these forms too have a single underlying morphological representation and a morphology-phonology mapping which is optionally unfaithful to the morphological structure in the case of mismatches.

The two approaches are only "mechanically" different. They both express the idea that morphological complexity has become optionally "obscured" for some native speaker(s) or that they sometimes "feel" a form consists of more than one domain, sometimes not – this is what parallel representations or optionally unfaithful mapping essentially mean. Both approaches can be criticised because neither the occurrence of parallel representations, nor unfaithful mapping (over- and underparsing), nor the optionality of unfaithfulness is connected in them with any other properties of the forms for which they are hypothesized and thereby predict that phonological vacillation resulting from mismatch is accidental and random. They make no prediction about which forms behave in this way, nor do they characterize the forms displaying

* This work has been supported by National Scientific Grant NKFI-119863 'Experimental and theoretical investigations of vowel harmony patterns'.

Proceedings of *AMP 2016*

phonological vacillation resulting from mismatch. Thus, in the traditional categorical and representational view it is arbitrary which forms behave in this way: they are the ones that happen to be represented as described above independently of their other properties.

In this paper we show, by examining the harmonic behaviour of Hungarian compounds, that phonological vacillation resulting from mismatch is related to and can be explicitly characterised with reference to the token frequencies of these forms and their parts. In section 2 we describe the relevant data. In section 3 we discuss Hay's measure of the parsability of derived forms. In section 4 we modify the Hayian measure, apply it to compounds and show how the harmonic vacillation of suffixed forms of some compounds can be characterised with the measure we propose. We conclude by summarising our analysis and identify some problems for further research.

2 The data

2.1 *The harmonic behaviour of compounds* Neutral vowels are transparent to backness harmony in Hungarian, i.e. the frontness/backness of a harmonising suffix is determined by the last non-neutral vowel of the stem (cf. e.g. Törkenczy 2011). The vowels **i**, **i**: **e**: and to a certain extent **ɛ** count as neutral. Thus, both monomorphemic [BN] roots and back roots with a neutral vowel suffix [[B]N] take back vowel alternants of harmonizing suffixes (e.g. **ɑli**:**z-nɑk** 'Alice-DAT', **hɑl-e**:**-nɑk** 'fish-POSS-DAT') while stems of the vocalic pattern [FN] and [[F]N] take front ones (e.g. **røvid-nɛk** 'short-DAT', **sɛm-e**:**-nɛk** 'eye-POSS-DAT').[1] However, in the case of a compound, it is not the vocalic pattern of the whole compound that determines its harmonic behaviour but only the vocalic pattern of its rightmost member. In some cases this can manifest itself as non-transparency. When the last member of a compound only contains neutral vowels, then the backness of the last harmonic vowel of the first member of the compound has no effect on the suffixes attached to the compound even though that vowel is the last non-neutral vowel of the base. Since all-neutral stems in **i**, **i**:, **e**: behave in two different ways (they may be harmonic, selecting front suffix alternants, e.g. **i**:**z-nɛk** 'taste-DAT', or antiharmonic, selecting back suffix alternants, e.g. **hi**:**d-nɑk** 'bridge-DAT'), there are two such cases: (i) if the first member is a harmonically *back* stem and the second member is a *harmonic* all-neutral stem, the suffix will be front and not back (the latter is expected after a BN pattern otherwise): [[B][N]$_{H}$]**F** e.g. **hɑl-i**:**z-nɛk** 'fish taste-DAT'; or (ii) if the first member is a harmonically *front* stem and the second member is an *antiharmonic* all-neutral stem, the suffix will be back: [[F][N]$_{AH}$]**B** e.g. **kø**:**-hi**:**d-nɑk** 'stone bridge-DAT'. Thus, compounds are harmonically different from other simplex or complex bases and in some contexts this results in near minimal pairs, e.g. **hɑl-e**:**-nɑk** vs. **hɑl-le**:**-nɛk** 'fish soup-DAT'; **sɛm-e**:**-nɛk** vs. **sɛm-he**:**j-nɑk** 'eyelid-DAT':

transparent neutral V		non-transparent neutral V
root internal	in suffix	in rightmost compound member
[BN]**B**	[[B]N]**B**	[[B][N]$_{H}$]**F**
[FN]**F**	[[F]N]**F**	[[F][N]$_{AH}$]**B**

Table 1 The transparency of neutral vowels in morphologically different contexts

The transparency of neutral vowels is limited in two ways. It is subject to the Height Effect, which applies to the neutral vowel **ɛ** and it is subject to the Count Effect, which applies to sequences of syllables with neutral vowels (cf. e.g. Hayes & Cziráky Londe 2006). In these cases, transparency is only partial, typically resulting in hesitation: [Bɛ]**B/F**, e.g. **fo**:**sɛr-nɑk/nɛk** 'guy-DAT', **kɑr-ɛs-nɑk/nɛk** 'Charles-DIM-DAT'; [BNN]**B/F**, e.g. **klɑrine**:**t-nɑk/nɛk** 'clarinet-DAT'. However, suffixed compounds of the same harmonic patterns show no variation. They always take a harmonically front suffix when the rightmost member is not antiharmonic: [[B][ɛ]]**F** (e.g. **po**:**t-sɛr-nɛk** 'ersatz substance-DAT', **fɑ-sɛs-nɛk** 'wood

1 We use the capital letters N, B, F to denote neutral, back and front vowels, respectively. Square brackets show morphological structure: monomorphemic root [XY], suffixed root [[X]Y], compound [[X][Y]]. These notations are just meant as shorthand and are not claims about morphophonological representations or any theoretical position.

alcohol-DAT'); [[BN][N]$_H$]**F** or [[B][NN]]**F** (pl. **hupi-ke:k-n<u>ɛ</u>k** 'cheap bright blue-DAT', **bor-vide:k-n<u>ɛ</u>k** 'wine_growing region-DAT'). These are summarised in Table 2.

transparent neutral V		non-transparent neutral V
root internal	in suffix	in rightmost compound member
[Bɛ]**B/F**	[[B]ɛ]**B/F**	[[B][ɛ]]**F**
[BNN]**B/F**	[[BN]N]**B** or [[[B]N]N]**B**	[[BN][N]$_H$]**F** or [[B][NN]]**F**

Table 2 Variation-inducing contexts and the lack of variation in compounds

2.2 *The problem: "obscured" compounds* Some words whose vocalic patterns are the same as those of the ones we summarised in Table 1 do not show categorical harmonic behaviour. The harmonic suffixes attached to them are subject to variation to some degree: **honve:d-n<u>ɑ</u>k/n<u>ɛ</u>k** 'soldier-DAT', **hu:ʃve:t-n<u>ɑ</u>k/%n<u>ɛ</u>k** 'Easter-DAT', **fe:rfi-n<u>ɑ</u>k/n<u>ɛ</u>k** 'man-DAT'. Thus, the neutral vowels **i**, **i:**, **e:** are not completely non-transparent in these words (so they are unlike compounds where the same vowels are non-transparent in the rightmost member), but they are not completely transparent either (so these words are unlike monomorphemic words or affixed roots where the same vowels are fully transparent). This means that the phonologically relevant morphological structure of these words does not fit into any of three possibilities we have identified (monomorphemic root, affixed stem, compound). **honve:d** and **hu:ʃve:t** are expected to be (a) categorically back harmonic as monomorphemic forms ([BN]**B**; like other such roots) or (b) categorically front harmonic as compounds ([[B][N]$_H$]**F**; like other compounds whose rightmost members are the front harmonic roots **ve:d** and **ve:t**, e.g. **ha:t-ve:d-n<u>ɛ</u>k** 'quarterback-DAT' and **uta:n-ve:t-n<u>ɛ</u>k** 'C.O.D.-DAT'). Similarly, **fe:rfi** should be (a) categorically front harmonic as a monomorphemic root or (b) categorically back harmonic as a compound ([[N][N]$_{AH}$]**B)** whose rightmost member is **fi** which is an antiharmonic root (**fi-<u>ɑ</u>m** 'son-1SG.POSS, **fi-<u>u</u>ŋk** 'son-1PL.POSS).

The problem in these cases is that the attested *variable* harmonic behaviour of these forms differs from the categorical harmonic behaviour which is predicted by their harmonic patterns (vocalic make-up) and/or their morphological structure. This is shown in Table 3:

root	compound	attested	example
[BN]**B**	[[B][N]$_H$]**F**	**B/F**	hon(-)ve:d-n**ɑ**k/n**ɛ**k hu:ʃ(-)ve:t-n**ɑ**k/%n**ɛ**k
[NN]**F**	[[N][N]$_{AH}$]**B**	**B/F**	fe:r(-)fi-n**ɑ**k/n**ɛ**k

Table 3 Variation vs. predicted categorical behaviour

Traditionally, these are called "obscured" compounds, but this term is a misnomer since compounds whose "compoundness" has been obscured should behave like simple forms of a single domain, i.e. they should show no variation in their harmonic behaviour. Assuming that harmonic variation is due to the fact that for some speakers they are still consistently compounds and for others they are consistently monomorphemic (obscured) is untenable because this would only explain interspeaker variation while here we find variation within the individual. Thus, using parallel representations cannot explain variation in these cases. Alternatively, we might think that speakers showing intraspeaker variation randomly assume a type of representation for these forms (a multi-domain compound-like one or a single-domain monomorphemic one). This view, however, is incompatible with the standard theories of lexical representations in representational models and has no advantage over a non-representational approach (in which behaviour is not encoded into representations). Furthermore, as we have pointed out in section 1, in these approaches the curious behaviour of these forms is seen as accidental since no connection is established between this behaviour and some other property of the relevant forms.

A related problem is associated with some forms whose harmonic pattern induces variation in monomorphemic roots because of the height effect or the count effect ([Bɛ], [BNN]). True compounds of the same harmonic patterns show no variation at all, cf. Table 2. While affixed forms of roots of this kind

show considerable variation, harmonically affixed "obscured compounds" show very little variation, which is typically substandard, e.g. **mo:dsɛr-nɛ̲k/%nɑ̲k** 'method-DAT', **ɟo:ɟsɛr-nɛ̲k/%nɑ̲k** 'medicine-DAT'. Other forms with the same right member show no variation at all (i.e. true compound behaviour), e.g. **ta:p-sɛr-nɛ̲k** 'food preparation-DAT', **po:t-sɛr-nɛ̲k** 'ersatz substance-DAT'. This is summarised in Table 4:

root [Bɛ]**F/B**	compound [[B][ɛ]]**F**	attested	examples
+	–	**F/B**	fo:sɛr-n**ɑ**k/n**ɛ**k ʃnɑpsɛr-n**ɑ**k/n**ɛ**k
?	?	**F/%B**	mo:d(-)sɛr-n**ɛ**k/%n**ɑ**k ɟo:ɟ(-)sɛr-n**ɛ**k/%n**ɑ**k
–	+	**F**	ta:p-sɛr-n**ɛ**k/*nɑk po:t-sɛr-n**ɛ**k/*nɑk

Table 4 Morphologically complex – phonologically ambiguous

BNN-forms show similar contrasts. The word **oksige:n** 'oxygen' can take (rare, substandard) back suffix alternants **oksige:n-nɛ̲k/%nɑ̲k** 'oxygen+DAT' while **ɑntige:n** 'antigen' can only take front ones: **ɑntige:n-nɛ̲k/*nɑk** 'antigen-DAT'. This ties in with the fact that both members of the latter (**ɑnti**, **ge:n**) occur in other compounds while the first member of the former (**oksi**) does not.

root [BNN]**F/B**	compound [[BN][N]]**F**	attested	examples
+	–	**F/B**	ɑbige:l-n**ɛ**k/n**ɑ**k
?	?	**F/%B**	oksi(-)ge:n-n**ɛ**k/%n**ɑ**k
–	+	**F**	ɑntige:n-n**ɛ**k/*nɑk

Table 5 Morphologically complex – phonologically ambiguous

The examples above in Table 3-5 show that the behaviour of morphologically complex forms is not categorical: some do not behave like compounds but do not behave like monomorphemic roots either.

Furthermore, there are forms that are certainly not compounds morphologically, but their harmonic behaviour is unlike those of monomorphemic roots. Consider the roots that end in the string **ne:z**, e.g. **mɑjone:z** 'mayonnaise', **mɛlɑne:z** 'Melanesian', **indone:z** 'Indonesian', **mikrone:z** 'Micronesian', **poline:z** 'Polynesian', **bɑline:z** 'Balinese', etc. These roots end in the harmonic patterns B**e:** or B**ie:** where the former type of roots *generally* shows near-categorical back harmonic behaviour and the latter is subject to a great degree of variation (cf. Rebrus & Törkenczy 2016b). By contrast, these particular words are predominantly front-harmonic in their behaviour (and show very little if any variation), e.g. **mɑjone:z-nɛ̲k/%nɑ̲k** 'mayonnaise+DAT', **poline:z-nɛ̲k/%nɑ̲k** 'Polynesian+DAT'. If we disregard the very small degree of variation, it is unclear how they can be front harmonic. This is only possible if they were compound words whose final vowel **e:** is in the second member of the compound. Note that this behaviour is not attested when the form does not end in the string **ne:z**: e.g. **siŋgɑle:z-nɛ̲k/nɑ̲k** 'Sinhalese+DAT', **trɑpe:z-nɛ̲k/nɑ̲k** 'trapezoid+DAT'. This is shown in Table 6.

root [B...N]**F/B**	compound [[...][N]]**F**	attested	examples
+	–	**F/B**	trɑpe:z-n**ɛ**k/n**ɑ**k siŋgɑle:z-n**ɛ**k/n**ɑ**k
?	?+	**F**/ʔB	mɑjo(-)ne:z-n**ɛ**k/%n**ɑ**k poli(-)ne:z-n**ɛ**k/%n**ɑ**k
–	+	**F**	odɑ-ne:z-n**ɛ**k/*nɑk sa:rgɑ-re:z-n**ɛ**k/*nɑk

Table 6 Morphologically simplex – phonologically complex

To sum up, we have seen that with some harmonic patterns the harmonic behaviour of compounds is different from that of monomorphemic roots. However, there are words that do not fit into either category on the basis of their harmonic behaviour, but are somewhere "in between" a compound and a non-compound. Informally, we can say that neither their monomorphemic status nor their compound status is particularly strong: they are semantically and/or morphologically not prototypical compounds. In what follows we argue that this half-obscured status can be quantitatively captured and characterised.

3 The parsability of derived words

In section 1 we discussed the traditional approach according to which morphological complexity is *idiosyncratic* information. Accordingly, it must be *stipulated* (for instance in the lexicon) if a form is complex or not and this information is not systematically related to or derivable from any other property of the given form. A different position is taken by other approaches (e.g. Bybee 1995) which rely on the observation that there is an inverse relationship between morphological complexity and *token frequency*: the more frequent a form is, the more likely it is that it is monomorphemic (or behaves like a single morpheme phonologically). In other words, frequency is a good indicator of unity, i.e. whether a form is stored as a unit in the lexicon. This idea is psychologically plausible since lexical access is frequency-sensitive: frequent forms are accessed faster because presumably they are accessed directly as wholes rather than being decomposed into their parts.

Hay (2001) and Hay & Baayen (2005) analyse the complexity of derived forms in a more sophisticated way: they propose to measure morphological complexity of a derived form with its *relative frequency* compared to the frequency of its base rather than its absolute frequency. The *parsability* or *decomposability* *d*(AB) of a derived form AB (where A is the base and B is the affix) is inversely proportional to the token frequency of AB and is directly proportional to the free occurrences of its base A.

(1) The parsability of derived forms:

$$d(\mathrm{AB}) = \mathrm{freq}(\mathrm{A}) / \mathrm{freq}(\mathrm{AB}) \qquad (0 \leq d < +\infty)$$

The measure *d* can be any nonnegative number. One limiting case is when the form under consideration is monomorphemic: in this case it cannot be parsed into A and B so that A occurs as a free form, i.e. the frequency of A (the numerator of the fraction above) is zero and therefore the parsability *d* of AB is zero independently of the frequency of AB. When A does occur in isolation but its frequency is low and/or the frequency of AB is high, *d* is very close to zero. This means that AB is not monomorphemic, but its parsability is very low: it is weakly decomposable. The other extreme case is when the frequency of the derived form AB is low and/or the frequency of its base in isolation is high. In this case parsability can be very high. Thus parsability is gradient, it can take any value between zero (the form is not decomposable) and any positive number.

(2) The morphological complexity of derived forms according to Hay

Parsability:	monomorphemic	less parsable	more parsable
Examples:	*in-ept* (*d*=0)	*in-sane* (*d*=0.58)	*in-firm* (*d*=26.48)

The English examples above show that even derived forms containing one and the same affix may be parsable to a different degree.[2] The form *inept* only "virtually" has the prefix *in-* since *-ept* never occurs as a free form – so its parsability *d*=0. By contrast, the forms *insane* and *infirm* both have bases that occur in isolation, but with very different frequencies: *sane* is much rarer as a free form than *firm*. Furthermore, the token frequencies of the prefixed forms are just the opposite: *insane* is more frequent than *infirm*. Consequently, the parsability of *infirm* is much higher than that of *insane*. The *d*-value of the former is 26.48 while the *d*-value of the latter is 0.58.

4 The morphological complexity of compounds

4.1 *A measure of parsability* We have to make some crucial modifications if we want to apply the measure introduced above to compounds. First, in compounds, we can take into consideration the free occurrences of *both* component members, not just one (the base) as in affixed words: we will measure the parsability of both members (left and right) and we will take their arithmetic mean to be the parsability of the whole compound.

The other modification we propose is the following: the formula in (1) only takes into consideration the free occurrences of the base A (freq(A)) of a derived (affixed) form AB. The occurrences of the same base with other affixes (AC, AD, etc.) are not counted. This information, however, is potentially important for parsability since some bases are bound: they only occur when affixed and never in isolation. According to (1) the parsability of these forms is zero since freq(A)=0. A similar state of affairs can occur when in a compound one member does not occur in isolation or is very infrequent. Therefore we suggest that not only the isolated occurrences of the members of a compound should be counted, but also their occurrences as members in other compounds, provided that they are in the same positions in these compounds as in the one whose parsability is being measured. We add up the token frequencies of all these forms and divide the sum with the token frequency of the compound examined. Accordingly, given a putative compound AB, if A occurs in isolation and in other compounds AC and AD, then the left parsability of the compound AB is obtained by adding up the token frequencies of these occurrences and dividing the sum by the token frequency of AB: (freq(A)+freq(AC)+freq(AD))/freq(AB). This is given as a general formula in (3a) where X can be any compound member including an empty string. The right parsability of a compound is calculated from the sum of the token frequencies of compounds of the form XB analogously, cf. (3b). The bilateral parsability of the compound is the arithmetic mean of the two values, cf. (3c).

(3) Parsability of compounds

a. left	$d_L(AB) = \Sigma_X freq(AX) / freq(AB)$	
b. right	$d_R(AB) = \Sigma_X freq(XB) / freq(AB)$	
c. bilateral	$d(AB) = (d_L(AB) + d_R(AB)) / 2$	$(1 \leq d < +\infty)$

Note that the minimum of parsability here is 1: if member A of a putative compound AB does not occur in any AX other than the compound AB itself, then left parsability according to (3a) is $d_L(AB)=freq(AB)/freq(AB)=1$. Similarly, right parsability is also $d_L(AB)=freq(AB)/freq(AB)=1$ if member B of the same compound AB does not occur in any XB other than the compound AB itself either (cf. 3b). When both the right and the left parsability of a compound takes the minimal value 1, then the bilateral parsability of the compound is also minimal $d(AB)=1$ according to (3c). In this case the form is not a compound. If, however, both left and right parsability or either is greater than 1, then the mathematical mean of their sum, the bilateral parsability of the form is also greater than 1, which means that the form is a compound (to some degree). Similarly to derived forms, the (bilateral) parsability of compounds has no

2 The *d*-values have been calculated on the basis of the frequencies given in Hay 2001, 2003.

upper bound, the number is only limited by the frequencies of the relevant forms in the corpus. The ratios in (3ab) may be different for different forms and therefore the parsability of compounds is gradient.

(4) The morphological complexity of compounds (compare figure (2))

Complexity:	non-compound	less parsable	more parsable
Examples:	somse:d	honve:d (3.13)	ha:t-ve:d (312.5)
	ɑŋke:t	hu:ʃ-ve:t (2.43)	uta:n-ve:t (227.3)
	profi, hifi	fe:r-fi (1.17*)	kira:j-fi (30.7*)

The examples (4) show that different compounds one of whose members is identical while the other is different may be radically differently parsable. For instance, the right member **ve:d** 'defence' in the words **hon-ve:d** and **ha:t-ve:d** is relatively infrequent in isolation or as the right member of compounds but **ha:t(-)** 'back' is much more frequent in isolation or as the left member of compounds than **hon(-)** 'home'. Consequently, and because **hon-ve:d** is much more frequent than **ha:t-ve:d**, the parsability of the former (2.70) is much lower than that of the latter (68.61). The same is true of the pairs **hu:ʃ-ve:t** (2.43) vs. **uta:n-ve:t** (227.3) and **fe:r-fi** (1.17) vs. **kira:j-fi** (30.7*) (frequency data are from *Szószablya* webcorpus, cf. Halácsy et al. 2004). The form **fe:r-fi** is especially interesting because its left member **fe:r-** only occurs in this word, but not in isolation or other compounds. Therefore the left parsability is minimal (d_L(**fe:r-fi**)=1). At the same time the right member **-fi** is also bound and occurs in compounds whose cumulative frequency is lower than the frequency of **fe:rfi** by one order of magnitude. The right parsability of **fe:r-fi** is not minimal but low (d_R(**fe:r-fi**)=1.09). As a consequence, bilateral parsability is low too (1.05). The parsability of non-compounds e.g. **somse:d** 'neighbour', **ɑŋke:t** 'conference', **profi** 'pro', **hifi** 'id', is minimal (d=1), i.e. they are not decomposable.

The parsability scale in (4) can be interpreted as an ordered sequence of the types *non-compound → obscured compound → true compound.* Forms whose parsability is greater than 1 but low fall into the traditional "obscured compound" category. In what follows we will try to identify the parsability range in which a form counts as an "obscured compound" and a "true compound". First, however, we have to take a closer look at the parsability measure in (3).

4.2 *Compound unity* We have seen that the parsability of compounds as defined by the formula in (3) is a value whose lower bound is 1, but which is not bounded from above, i.e. the value is a number in the interval [1,+∞). Such a measure is unwieldy; it would be much better to have one with both a lower and an upper bound. It is possible to derive such a measure from d by taking its reciprocal because the reciprocal of a number greater than 1 is a (positive) number smaller than 1. Thus, if d is in the interval [1,+∞), then $1/d$ is in the interval (0,1]. Since the reciprocal function $y=1/x$ is a strictly monotonically decreasing one, this new value measures the degree to which a compound is *unparsable* rather than its *parsability* (decomposability). We will refer to it as *morphological unity* (u). We present the formula in (5):

(5) The unity of compounds:

$$u(\mathrm{AB}) = 1 / d(\mathrm{AB}) \qquad (0<u\leq 1)$$

When unity is maximal (u(AB)=1), then d is minimal (d(AB)=1) and left and right parsability are also minimal and therefore the form being examined is not a compound. When, however, u is close to 0 (it can never be 0)[3], this means that d is relatively high and the form is highly parsable. These are what we consider true compounds traditionally, e.g. **ha:t-ve:d**, **uta:n-ve:t**, **kira:j-fi**. For those forms whose unity u is considerably higher than zero but does not reach the maximum of 1, parsability d is low and their compound status is uncertain, it has become "obscured", e.g. **honve:d**, **hu:ʃ-ve:t**, **fe:r-fi**. Those forms whose unity is maximal u=1 are not compounds, e.g. **somse:d**, **ɑŋke:t**, **profi**.

3 Unity u can be taken to be zero in a very special case when a compound AB does not occur in the corpus but its members do (at least in other compounds). In this case both left and right unity are 0 according to the formula in (7ab) since freq(AB)=0 and consequently bilateral unity is also 0 according to (7c).

(6) The morphological complexity of compounds (compare figure (4))

Complexity:	low unity	intermediate unity	maximal unity
Examples:	ha:t-ve:d (0.003)	honve:d (0.32)	somse:d
	uta:n-ve:t (0.004)	hu:ʃ-ve:t (0.41)	ɑŋke:t
	kira:j-fi (0.033*)	fe:r-fi (0.86*)	profi, hifi

Given the definition of unity in (5), we can define *left and right unity* as the reciprocal of left and right parsability (3ab), respectively; this is shown in (7ab) below. It is clear from the definition of bilateral parsability (3c) that bilateral unity is the reciprocal of the arithmetic mean of the reciprocal of left and right unity, i.e the harmonic mean of left and right unity $H(u_L,u_R)$. This is shown in (7).

(7) The unity of compounds

a. left $u_L(AB) = freq(AB) / \Sigma_X freq(AX)$
b. right $u_R(AB) = freq(AB) / \Sigma_X freq(XB)$
c. bilateral $u(AB) = H(u_L(AB), u_R(AB)) = 2 \cdot freq(AB) / (\Sigma_X freq(AX)+\Sigma_X freq(XB))$

It is an important property of the harmonic mean that the harmonic mean of a number close to zero and any number is a number close to zero. If we take the harmonic mean of 1 (the maximal value in our case) and a number close to zero, the result is a very small number, e.g. $H(0.01,1)\approx 0.02$. For compounds this means that a form whose unity is maximal 1 on one side (it is monomorphemic) but close to zero on the other side (a "true" compound) will be a "true" compound since its bilateral unity will be close to zero. This predicts that compounds which contain a cranberry morph will be evaluated as true compounds, provided that the other member frequently occurs in other compounds and/or in isolation. In other words, the parsability to a sufficient degree of *one* member of a compound makes the compound (bilaterally) parsable to a sufficient degree – even when its other member is not parsable at all.

4.3 *Unity and harmony* The measure of unity makes it possible to characterise compounds so that morphological complexity/unity is gradient rather than categorical and is grounded in language use (frequency). In order to determine how plausible the proposed measure is we need to connect the measure of morphological unity proposed with some phenomenon which is observable independently of frequency. We do this by comparing unity as defined in (7) with harmonic behaviour.

Let us examine the harmonically problematic forms we discussed in section 2.2. The unity values of the forms that show unexpected harmonic behaviour) are the following: u(**hon-ve**:**d**)=0.32, u(**hu**:**ʃ-ve**:**t**)=0.41, u(**fe**:**r-fi**)=0.86, i.e they are positioned around the middle of the unity scale ranging from 0 to 1. If we compare these values with those of compounds that have the same second members but are well-behaved harmonically, we find that the unity values of the latter are very close to zero: u(**ha**:**t-ve**:**d**)=0.003, u(**uta**:**n-ve**:**t**)=0.004, u(**kira**:**j-fi**)=0.03. Roots that happen to end in the same strings (e.g. **somse**:**d**, **ɑŋke**:**t**, **hifi**) naturally have maximal unity u=1. Tables 7-8 below show this together with statistical data about the harmonic behaviour of these items based on the Szószablya webcorpus. The last four columns of Tables 7-8 show the F-ratio measured in word tokens and types of affixed forms and the number of relevant types and tokens (the F-ratio is the ratio of the number of front suffixed forms compared to the number of all harmonically suffixed forms: 100% is categorical front harmonic behaviour (**F**) and 0 is categorical back harmonic behaviour (**B**) cf. Rebrus & Törkenczy 2016b).

Table 7 shows that the F-ratio of **ha**:**tve**:**d** is 100%, i.e. it only takes front harmonic suffixes. Harmonically, **somse**:**d** is (almost) always back because **e**: is transparent: F-ratio is 0.01% in tokens and 8% in types. By contrast, **honve**:**d** shows a high degree of variation: its F-ratio is 42.77% in tokens and 55% in types. Similarly, categorical vs. variable behaviour occurs with compounds whose right member is antiharmonic (e.g. **kira**:**j-fi** vs. **fe**:**r**(**-**)**fi**), see Table 8. The unity of true compounds is low and they behave (near) categorically (the F-ratio of **kira**:**j-fi** is 0.45% in tokens, 5% in types. Monomorphemic **hifi** is

categorically front (F-ratio=100%). By contrast, the "obscured" compound **fe:rfi** shows a high degree of variation, its F-ratio is 20.14% in tokens and 48% in types.[4]

word	*u*	harmony	F-ratio (token)	N (token)	F-ratio (type)	N (type)
ha:t-ve:d	0.003	front	100.00%	183	100%	18
hon(-)ve:d	**0.320**	back~front	**42.77%**	807	**55%**	31
somse:d	1.000	back	0.01%	13366	8%	26

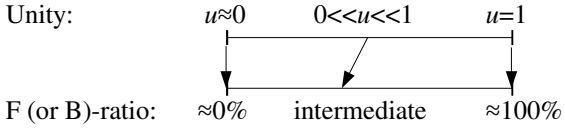
Table 7 Unity and harmony (webcorpus)

word	*u*	harmony	F-ratio (token)	N (token)	F-ratio (type)	N (type)
kira:j-fi	0.033	mainly back	0.45%	889	5%	19
fe:r(-)fi	**0.855**	back~front	**20.14%***	14353*	**48%***	27*
hifi	1.000	front	100.00%	129	100%	13

Table 8 Unity and harmony (webcorpus)

The data above show that there is a connection between the "intermediate" morphological unity of forms and their non-categorical harmonic behaviour: the F-ratios of forms whose unity differs considerably from the extremal values 1 and zero are also significantly different from 0% and 100%. This means that the forms that are intermediate on the scale ranging from true compounds to monomorphemic roots are also intermediate on the scale ranging from categorical front to categorical back harmonic behaviour, i.e. "obscured compounds" show hesitation in harmony. This is shown in (8) where arrows indicate that the extrema and the intermediate values of the two scales are mapped onto one another.

(8) Harmonic variation with near-categorical regular harmonic scalar extrema



The exact value of the F ratio (or its difference from 100%, the B ratio) depends on other factors, too, e.g. the quality of the neutral vowels (**i/i:** vs. **e:**), the number and quality of stem-final consonants etc. (cf. Hayes *et al.* 2008, Rebrus & Törkenczy 2016b). Therefore, we cannot claim that there is a correlation between the measure of unity *u* and F-ratio until we can quantify the contribution of these other effects. What we can claim now is that a certain deviation from the extremal *u* values 0 and 1 (the unity of a true compound vs. a monomorphemic root) co-occurs with a some amount of hesitation in cases when generally the harmonic behaviour of a true compound vs. a monomorphemic root is categorical and different.

Next, we will examine the type of forms which show categorical behaviour at only one extremum of the complexity scale (the true-compound end) but display a high degree of vacillation at the other end, when their base is monomorphemic (B**ɛ** type and BNN type, cf, Table 3-4. The unity values of words like **po:tsɛr** and **ta:psɛr**, which count as "true" compounds are close to zero and their harmonic behaviour is almost categorical (front): their B-ratios[5] are under 0.01%, see (13) below. Forms based on the monomorphemic items of this type (e.g. **fo:sɛr**, **ʃnɑpsɛr** <card game>) show a high degree of variation:

4 In Table 8 we only counted forms with consonant-initial suffixes because two highly frequent forms (**fe:rfi-ɑk** 'man-PL', **fe:rfi-ɑʃ** 'man-ly') only occur with back suffix alternants, and including them would skew the statistics.

5 There are two differences in Table 9 compared to the previous two tables 7-8: (i) B-ratio is shown instead of F-ratio; (ii) the numbers are based on a Google search of forms with specific suffixes (instrumental **-vɑl/vɛl** and dative **-nɑk/nɛk**) because the small degree of substandard variation would have been undetectable in webcorpus.

the relevant B-ratios are **foːsɛr**: 48.51% / 47.70%; **ʃnɑpsɛr**: 9.43% / 4.39%. The words **ɟoːɟsɛr** and **moːdsɛr** have *u*-values which are well above zero (*u*=0.08, ill. *u*=0.18, respectively) and consequently show some degree of variation in accordance with what has been said above. This manifests itself in their B-ratios (0.14% / 0.12% and 0.17% / 0.18%, respectively), which are considerably higher than the B-ratio of true compounds (which is lower than 0.01%).

word	***u***	**harmony**	**B-ratio *-vɑl/vɛl***	**N *-vɑl/vɛl***	**B-ratio *-nɑk/nɛk***	**N *-nɑk/nɛk***
poːt-sɛr	0.0006	categorical:	0.00%	1390	0.00%	1620
taːp-sɛr	0.004	front	0.006%	31900	0.00%	15000
ɟoːɟ(-)sɛr	**0.078**	low degree of variation	**0.14%**	207300	**0.12%**	124100
moːd(-)sɛr	**0.175**	front(~back)	**0.17%**	1572600	**0.18%**	215400
foːsɛr	1.000	high degree of variation:	48.51%	7710	47.70%	11500
ʃnɑpsɛr	1.000	front~back	9.43%	106	4.39%	139

Table 9 Unity and harmony (Google search)

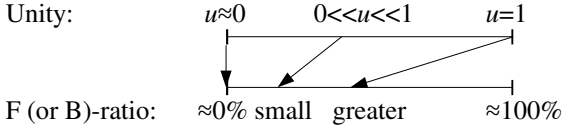
Forms like **oksigeːn** and **ɑntigeːn** behave similarly. The unity of **ɑntigeːn** is relatively low (*u*=0.02) since it is infrequent while the frequency of its first member **ɑnti** is high. **oksigeːn** is more frequent and its first member **oksi** is unattested in isolation and in other compounds, consequently its unity is intermediate (*u*=0.49). Harmonically **ɑntigeːn** behaves like a true compound, i.e its harmonic behaviour is determined by that of the rightmost member, i.e. it only takes front suffixes. **oksigeːn** is different: its B-ratio measured in tokens is low (0.36%) but a significant number of the harmonic suffixes that it cooccurs with are back (23%). Similar monomorphemic roots (in accordance with the general behaviour of BNN roots) show a high degree of hesitation, e.g. the B-ratios of **ɑːbigeːl** 'Abigel' are 7.33% (tokens) and 29% (types).

word	***u***	**harmony**	**B-ratio (token)**	**N (token)**	**B-ratio (type)**	**N (type)**
ɑnti-geːn	0.016	front	0.00%	134	0%	8
oksi(-)geːn	**0.493**	front(~back)	**0.36%**	1665	**23%**	22
ɑːbigeːl	1.000	front~back	7.32%	41	29%	7

Table 10 Unity and harmony (webcorpus)

Summing up the last two cases we can see that here too the harmonic behaviour of the forms of intermediate unity is intermediate between the harmonic behaviour of two extremal cases of unity, see figure (8). The difference is that in Table 9-10 the harmonic behaviour of monomorphemic bases is not categorical. The harmonic behaviour of "obscured" compounds is consequently not simply variable, but variable to a small degree, here it is close to categorical front harmonicity. This hesitation may be so small that it might be missing from some varieties of Hungarian (this is why we call it substandard variation).

(9) Variation with near-categorical and non-categorical regular harmonic scalar extrema



5 Conclusions and further research

Table (16) summarises the harmonic behaviour of "obscured compounds" (row ii.) compared to that of true compounds (row i.) and monomorphemic roots (row iii.) together with their unity values shown in the first column. The relevant kinds of harmonic behaviour are labelled in each cell at the intersections of rows i., ii. and iii. and columns a., b. and c. and are also given symbolically as α and β, where α and β are variables each representing the suffixal harmonic values F or B. α and β must take non-identical values when they both appear in the same cell or in the same column. The last row provides "obscured" compounds exemplifying the type of harmonic behaviour that occurs in row ii. in the relevant column. We can see in column (16a) that "obscured" compounds (whose unity is between 0 and 1) show hesitation in affix harmony when true compounds and roots of the same phonological make-up behave categorically and *differently from each other* in palatal harmony (cf. 3a). Column (16b) shows that when the phonological makeup of a form is such that it displays categorical behaviour in affix harmony as a true compound, but considerable variation as a monomorphemic root, then obscured compounds are different from both in that they do have variable affix harmony, but to a significantly smaller degree than roots. Column (16c) shows the case when the phonological makeup of a form happens to be such (e.g. Fɛ) that it behaves in the same way harmonically (categorically) irrespective of its morphological complexity, its *u* value.

Thus, "*prototypical*" harmonic behaviour occurs when (i) *u* is 1: single-domain behaviour; or (ii) when *u*=0 (multiple-domain behaviour). "*Unusual*" harmonic behaviour (which manifests itself (a) in variation instead of categorical behaviour (16ab) or (b) a degree of variation lower than expected (16b) occurs between a bottom and a top threshold in unity.

unity	**morphological status**	**harmonic behaviour of extremal types**		
		a. opposite	**b. partially opposite**	**c. identical**
$u \approx 0$	i. true compound	categorical (α)	categorical (α)	categorical (α)
$0 << u << 1$	ii. "obscured"	hesitation (α~β)	small hesitation (α~$^{\%}$β)	categorical (α)
$u = 1$	iii. non-compound	categorical (β)	greater hesitation (α~β)	categorical (α)

Table 11 The harmonic behaviour of "obscured compounds"

The harmonic behaviour of compound(-like) forms is thus related to their complexity, which can be characterised with the measure of unity *u*. It is clear, however, that while the frequency-based measure *u* as defined here is in correspondence with the morphological complexity of compounds, the parsability of compounds is more complex and depends also on factors not considered here, e.g. phonotactics, semantic transparency/compositionality, etc. The possible refinements of the measure of unity (e.g. by weighted mean) and the relationship between a frequency-based measure and other conceivable indicators/factors of complexity are issues for further research. These are intricate questions and the incorporation of such other aspects in an appropriate measure of unity is not straightforward. We mention here one semantics-related problem to illustrate this. We noted earlier that we do not count coincidences of substrings when the string identities do not correspond to semantic/morphological identity. However, the data in Table 6, when morphologically simplex forms appear to be treated harmonically as if they were intermediate between simplex and complex (e.g. **poli(-)ne:z-nɛk/$^{\%}$nɑk, mɑjo(-)ne:z-nɛk/$^{\%}$nɑk**, etc.), are problematic in this respect because this kind of "misanalysis" (the lowering of unity) only occurs when the meaningless final sequence is recurrent and frequent as some other semantically and morphologically completely unrelated form (in this case the extremely frequent front harmonic morpheme **ne:z** 'look'). Thus this behaviour is motivated by frequency, but seems insensitive to semantic/morphological identity: the former aspect is central while the latter is contrary to what our measure of unity generally assumes.

References

Bybee, Joan. (1995) Diachronic and typological properties of morphology and their implications for representation. In *Morphological Aspects of Language Processing*, ed. by Laurie Beth Feldman, 225–246. Hillsdale, NJ: Erlbaum.

Chomsky, Noam and Morris Halle. (1968) *The Sound Pattern of English.* New York: Harper & Row.

Halácsy, Péter, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón (2004) Creating open language resources for Hungarian In: *Proceedings of Language Resources and Evaluation Conference (LREC04)*. LREC, 203–210; http://szotar.mokk.bme.hu/szoszablya/searchq.php

Hay, Jennifer B. (2001) Lexical frequency in morphology: is everything relative? *Linguistics* 39. 1041-1070.

Hay, Jennifer B. (2003) *Causes and Consequences of Word Structure*. New York & London: Routledge.

Hay, Jennifer B. and Baayen, R. Harald. (2005) Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342-348.

Hayes, Bruce and Zsuzsa Cziráky Londe (2006) Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23. 59–104.

Hayes, Bruce, Kie Zuraw, Péter Siptár, and Zsuzsa Londe. (2009) Natural and unnatural constraints in Hungarian vowel harmony. *Language* 85. 822–863.

Kiparsky, Paul (1982) From Cyclic to Lexical Phonology. In *The structure of phonological representations, vol. I*, ed. by Harry van der Hulst and Norval Smith, 131–75. Dordrecht: Foris Publications.

McCarthy, John. (1981) A prosodic Theory of Nonconcatenative Morphology, *Linguistic Inquiry* 12. 373-418.

Rebrus, Péter and Miklós Törkenczy. (2015) Monotonicity and the typology of front/back harmony. *Theoretical Linguistics* 41/1–2. 1–61

Rebrus, Péter and Miklós Törkenczy. (2016a) Monotonicity and the limits of disharmony. In *Proceedings of the 2014 Annual Meeting on Phonology*, ed. by Adam Albright and Michelle A. Fullwood. Washington, DC.: Linguistic Society of America.

Rebrus, Péter and Miklós Törkenczy. (2016b) A Non-cumulative Pattern in Vowel Harmony: a Frequency-Based Account. In *Proceedings of the 2015 Annual Meeting on Phonology*, ed. by Gunnar Ólafur Hansson, Ashley Farris-Trimble, Kevin McMullin and Douglas Pulleyblank. Washington, DC.: Linguistic Society of America.

Rebrus, Péter and Miklós Törkenczy. (2017) Co-patterns, subpatterns and conflicting generalizations in Hungarian vowel harmony. In *Approaches to Hungarian. Volume 15: Papers from the 2015 Leiden Conference*, ed. by Harry van der Hulst and Anikó Lipták, Amsterdam/Philadelphia: John Benjamins.

Rose, Sharon and Rachel Walker. (2011) Harmony Systems. In *Handbook of Phonological Theory.* 2nd ed., ed. by John Goldsmith, Jason Riggle and Alan Yu, Cambridge, MA: Blackwell. 240-290.

Törkenczy, Miklós. (2011). Hungarian vowel harmony. In *The Blackwell Companion to Phonology,* ed. By M. van Oostendorp, C. J. Ewen, E. Hume and K. Rice, 2963–2990. Malden, MA & Oxford: Wiley-Blackwell.